SISTEMA DE RECONHECIMENTO DE VOZ PARA GERAÇÃO DE TEXTO PARA AUXILIAR INDIVÍDUOS COM DEFICIÊNCIA

Valber Antônio Gonçalves, Luciene Chagas de Oliveira
Universidade de Uberaba - Uniube, Campos Uberlândia - MG
valber.antonio@gmail.com, lchagasoliveira@gmail.com

Resumo **– A tecnologia tem sido cada dia um grande aliado e facilitador, que proporciona à pessoa com deficiência maior independência, qualidade de vida e inclusão social, através da ampliação de sua comunicação, mobilidade, controle de seu ambiente, habilidades de seu aprendizado, trabalho e integração com a família, amigos e sociedade. Por isso é importante na área de pesquisa de soluções para este grupo de pessoas. O objetivo deste trabalho é propor um sistema de reconhecimento de voz para geração de texto utilizando software livre para auxiliar indivíduos com deficiência com diversas tarefas diárias através de comandos extraídos dos textos gerado pelo sistema.**

Palavras-Chave **- Reconhecimento de voz, Geração de texto, decodificação de comando, deficientes.**

SPEECH RECOGNITION SYSTEM FOR TEXT GENERATION TO ASSIST INDIVIDUALS WITH DEFICIENCY

Abstract **- The technology has been every day a great ally and facilitator who provides the person with disabilities greater independence, quality of life and social inclusion through the expansion of its communication, mobility, your environment control, your learning skills, work and integration with family, friends and society. So it is important in research solutions for this group of people. The objective of this work is to propose a speech recognition system for generating text using free software to assist individuals with disabilities with various daily tasks through commands extracted from the texts generated by the system.**

Keywords **- voice recognition, text generation, command decoding, deficient.**

I. INTRODUÇÃO

Deficientes são pessoas que apresentam necessidades próprias e diferentes que requerem atenção específica em virtude de sua condição de deficiência. Genericamente também são chamados de portadores de necessidades especiais. Apresentam significativas diferenças físicas, sensoriais ou intelectuais, decorrentes de fatores inatos ou adquiridos, de caráter permanente, que acarretam dificuldades em sua interação com o meio físico e social.

A deficiência física ou motora é uma variedade de condições não sensoriais que afetam o indivíduo em termos de mobilidade, coordenação motora geral.

Essas pessoas necessitam de uma atenção especial, portanto faz-se necessário a existências de pesquisas e desenvolvimento de projetos tecnológicos com foco neste grupo. [1]

A tecnologia pode proporcionar uma melhor qualidade de vidas para as pessoas com deficiência.

A tecnologia assistiva é um termo ainda novo. Utilizado para identificar todo o arsenal de recursos e serviços que contribuem para proporcionar ou ampliar habilidades funcionais de pessoas com deficiência e consequentemente promover vida independente e inclusão.

Em meio à quantidade de deficientes existentes, especialmente aqueles com deficiência motora, que possuem dificuldades de locomoção, vê-se a necessidade de aplicar recursos visando melhorar a inclusão dessas pessoas. [2]

Tecnologias de reconhecimento da fala (também denominado em alguns aparelhos como reconhecimento de voz) permitem que computadores equipados com microfones interpretem a fala humana, por exemplo, para transcrição ou como método de comando por voz. Tais sistemas podem ser classificados por requererem, ou não, que o usuário treine o sistema a reconhecer seus padrões particulares de fala, por ter a habilidade de reconhecer fala contínua ou por requerer que o usuário fale pausadamente, e pelo tamanho do vocabulário que é capaz de reconhecer (pequeno, da ordem de dezenas a centenas de palavras, ou grande, com milhares de palavras).

Utilizar este recurso é a ideia inicial para o desenvolvimento de uma hardware capaz de identificar os comandos falados pelo usuário e transformar em sinais elétricos que poderá ser utilizado para acionamento de equipamentos que o mesmo utiliza.

II. SÍNTESE DE VOZ

Síntese de voz é o processo de produção artificial de voz humana. Um sistema informático utilizado para este propósito é denominado sintetizador de voz, e pode ser implementado em software ou hardware. Um sistema texto para voz (ou TTS em inglês) converte texto em linguagem normal para voz; outros sistemas interpretam representação linguística simbólica (como transcrição fonética) em voz. Voz sintetizada pode ser criada concatenando-se pedaços de

XIV CEEL - ISSN 2178-8308
03 a 07 de Outubro de 2016
Universidade Federal de Uberlândia - UFU
Uberlândia - Minas Gerais - Brasil

fala gravada, armazenada num banco de dados. Os sistemas diferem no tamanho das unidades de fala armazenadas; um sistema que armazene fones ou alofones fornecem a maior faixa de saída, mas podem carecer de clareza. Para usos específicos, o armazenamento de palavras ou frases inteiras possibilita uma saída de alta qualidade. Alternativamente, um sintetizador pode incorporar um modelo do trato vocal (caminho percorrido pela voz) e outras características da voz humana, para criar como saída uma voz completamente "sintética". [3]

A qualidade de um sintetizador de voz é determinada por sua similaridade com a voz humana e por sua capacidade de ser entendida. Um programa TTS inteligível permite que pessoas com deficiência visual ou com problemas de leitura possam ouvir obras escritas num computador pessoal. Muitos sistemas operacionais têm incluído capacidade de síntese de voz desde o início da década de 1980. Várias aplicações e equipamentos utilizam este recurso para auxiliar e facilitar o dia a dia não só de um deficiente, mas também aplica à todos. Hoje é comum a utilização destes tipos de sistemas baseados em TTS. Operadoras de telefonia já substituem pessoas por sistemas que voz artificiais, através destes recursos transformam roteiros escritos em voz. O desafio para este tipo de sistema é produzir a voz de forma mais semelhante possível do ser humano “fluência”. Esse tipo de sistema também está presente em GPS para navegação que detalha a orientação por voz durante a navegação ou seja consegue dizer o nome das ruas ao percorrer a rota calculada pelo GPS. [4]

III. RECONHECIMENTO DE VOZ

Programas de atendimento eletrônico fazem parte dessa tecnologia. Eles ouvem o som emitido pela sua voz, classificam as sílabas e é aplicado um método de busca para associar estas informações com padrões de palavras a fim de encontrar semelhanças. Um aplicativo relativamente comum para celular e que tem feito sucesso entre os usuários e que também pode ser enquadrado nesta categoria é aquele que ao “ouvir” determinada música faz a identificação dela e de seu autor.

A partir disso, você já deve ter se lembrado de algumas outras aplicações comuns para o reconhecimento de voz. Para que o computador reconheça o som da sua voz juntamente com a fonética da palavra pronunciada e efetue a aplicação desejada, ele precisa encadear uma sequência de passos, primeiro ele precisa digitalizar a fala que se quer reconhecer. A voz humana consiste no som produzido pelo ser humano usando suas cordas vocais para falar, cantar, gritar, etc. Sua frequência varia entre 50 e 3400 Hz. Para reconhecer a fala, o computado utiliza um conversor analógico-digital que capta as vibrações criadas pela voz e converte essas ondas em dados digitais. [5]

Em seguida, aplica-se uma medida para cada uma das ondas captadas e o som digitalizado é filtrado para separá-lo de ruídos e interferências. Então, efetua-se uma computação das características que representam o domínio espectral (frequências) contido na voz. Nessa etapa do processo, o som pode necessitar ser sincronizado, pois as pessoas não costumam utilizar o mesmo tom e nem sempre falam na mesma velocidade. Isso consiste em um ajuste com modelos de som já armazenados na memória do classificador. Então essa digitalização é separada em frações ainda menores, ou seja, sons fonéticos não maiores do que uma sílaba. Em seguida, o programa compara os sons captados com fonemas conhecidos e presentes em seu banco de dados que correspondam ao idioma que o locutor tenha falado. Em outras palavras, é aplicado um método de busca para associar as saídas com padrões de palavras e da voz de quem as emitiu. [6]

Por último, o sistema analisa o resultado e o compara com palavras e frases conhecidas e, como resultado, ele identifica o que seu usuário disse e converte para a funcionalidade desejada (texto em uma planilha, um comando, o reconhecimento do usuário, etc...). Os sistemas que possuem um número de palavras mais limitado em seu banco de dados são aqueles aplicados para utilização por um grande número de usuários. Este calculador é programado para ser mais generalizado e embora haja variedade na fala (como diferentes tons de voz, sotaques, etc.) ele tem uma grande capacidade de reconhecimento. Um bom exemplo disso são os atendimentos eletrônicos que são realizados por operadoras de telefone.

DIFICULDADES E PONTOS FRACOS

Algumas das complicações mais conhecidas com relação aos sistemas de reconhecimento de voz são, por exemplo, com relação à fala contínua. Para o cérebro humano é fácil ouvir uma frase e rapidamente fazer a junção entre as palavras. Já para um computador, ele compreende mais facilmente se cada palavra for pronunciada separada e pausadamente em uma frase.

Há alguns anos, esse era um grande tabu para os programas que faziam reconhecimento, embora ainda hoje esse problema ainda possa ocorrer em determinados sistema. Outro problema relacionado a isso é a pronuncia. Por exemplo, a frase correta seria “como você está?”, porém a maioria das pessoas falaria “como ‘cê tá?”. Para um classificador isso representa uma grande diferença e ele pode acabar por entender uma frase completamente diferente do que aquela que foi pronunciada. Regionalismos, como sotaques e dialetos também podem alterar bastante a maneira como certas palavras ou frases são faladas e, portanto, a interpretação do sistema. [7]

Outra dificuldade encontrada por estes sistemas é a de separar falas simultâneas de vários usuários. Quando utilizados para “transcrever uma reunião”, por exemplo, eles podem ter problemas na identificação de palavras que foram sobrepostas porque duas pessoas falaram algo ao mesmo tempo, ruído e interferências pode confundir ou invalidar o reconhecimento e também decodificar fala com música de fundo é problema para esse tipo de sistema.

Isso acontece porque o programa precisa "ouvir" as palavras faladas de maneira correta para que possa diferenciá-las. Muitas falas ao mesmo tempo (ou excesso de barulho) podem fazer com que ele perca o foco e não funcione corretamente.

Homônimos (como "conserto" e "concerto") representam outro problema para alguns classificadores, pois não é possível diferenciar tais palavras baseando-se apenas no som. Pensando nisso, alguns já foram treinados para levar em consideração o contexto da palavra, resolvendo tal problema.

O classificador precisa fazer o tratamento de frases dividindo-as em palavras. Para isso ele precisa reconhecer em que momento cada palavra começa e termina. A partir disso ele cria a cadeia dos sons enfileirando os fonemas, montando palavras e construindo as frases. Se as vezes até nós não conseguimos compreender uma frase vinda de outra pessoa, imagine uma máquina. [6]

APLICAÇÕES

Existem várias aplicações comuns para esta tecnologia. Além das já citadas durante o texto (atendimento eletrônico, conversão de fala em texto, reconhecimento de música/autor e de pessoas), ainda existem as que são responsáveis por uma série de aplicações para telefones celulares, como a discagem falada (que reconhece o nome armazenado na memória e inicia a ligação), por exemplo.

Outra aplicação ao reconhecimento de sons que tem sido utilizada é a de identificação de disparos. Para ela, é utilizado um programa que reconhece de onde veio um tiro, a partir de uma série de captadores que recebem o som. Em seguida ele analisa o tipo de ruído para diferenciá-lo de fogos de artifício, morteiros, rojões, etc. E, por último ele calcula dados como velocidade e intensidade do som para poder calcular as coordenadas da emissão do disparo.
Programas para acessibilidade relacionados à fala e reconhecimento dela também utilizam tal tecnologia para suas funções como legendas oculta em canais de televisão. [8]

IV. DESENVOLVIMENTO DO PROJETO

Diferente do sistema apresentado anteriormente síntese de voz, a proposta deste trabalho é realizar o contrário que é capturar e armazenar a voz depois submeter esse áudio em um tratamento de reconhecimento, depois retornar a fala em forma de texto e disponibilizá-la em um hardware que seja prático e multifuncional, tanto para auxiliar portadores de deficiência quanto realizar outras tarefas especificadas.
O projeto consiste em um software para conversão de voz que transforma em texto para ser transmitido ao hardware.
Na figura 1 mostra o ambiente web criado para o reconhecimento da voz através de um microfone em que o usuário necessita apenas em colocar o idioma.

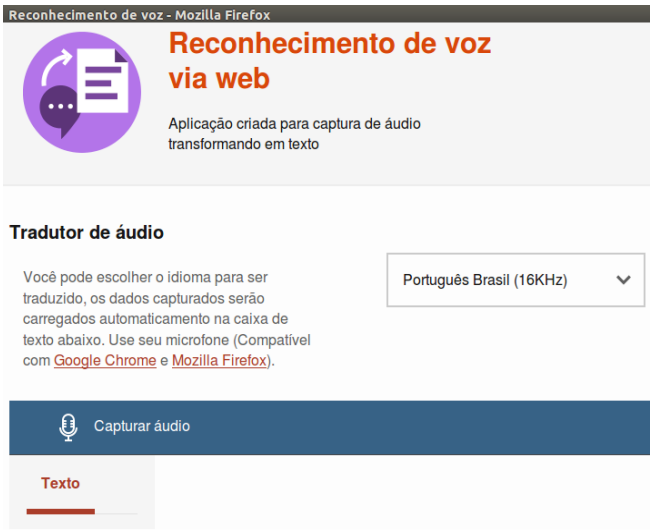

Fig. 1. Aplicação web para o reconhecimento de áudio.

Aqui pode-se verificar as características básicas para se construir uma aplicação que “ouve” o usuário e realiza uma busca pelo que foi dito. Após a análise da voz, o texto é transcrito para uma caixa de texto, conseguir reconhecer e atribuir uma ação através da fala abre inúmeras possibilidades de desenvolvimento! O processo de desenvolvimento de software aqui apresentado sugere uma metodologia de desenvolvimento de software de reconhecimento de voz, utilizando um sistema de computação em nuvem fornecido pela IBM "Bluemix" disponibilizado pelo conselho acadêmico da Universidade de Uberaba departamento de engenharia da computação.

Este estudo está inserido em um projeto de pesquisa que necessita de uma solução para a área de reconhecimento de voz. A motivação para o desenvolvimento de um software de reconhecimento de voz utilizando a arquitetura computacional da IBM Bluemix, surgiu da necessidade de captar a voz e reprocessá-la transformando em texto e posteriormente conduzi-lo a um hardware que transformará os comandos decodificados dos textos em sinal elétrico que poderá ser também contatos NA/NF para acionamento de cargas, que de forma geral para auxiliar na inserção de pessoas com deficiência realizar atividades significativa.

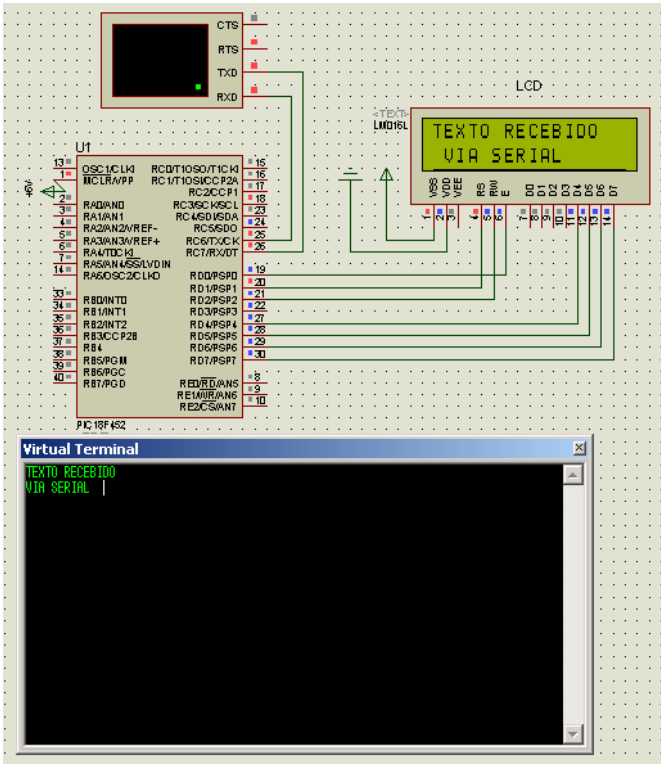

Fig. 2. Simulação do hardware.

Outra forma seria enviar os textos gerados através da captação do áudio de uma programação de televisão e inserir de forma serial num encoder de processamento de vídeo para inserir os caracteres nas linhas do vídeo. As emissoras de televisão preocupa-se em transmitir a todas as pessoas o conteúdo exibido na programação e, para isso, mantém o recurso de acessibilidade, o closed caption. O recurso de closed caption, ou legenda oculta, são as legendas com as transcrições das falas dos personagens e dos ruídos sonoros presentes nos programas de TV, séries, filmes e desenhos animados. O telespectador que deseja acompanhar a programação das emissoras de TV, canais legislativos ou televisão corporativa, que possui o recurso disponível com closed caption nos canais. Basta ativar o closed caption, o telespectador deve ativar a tecla CC do controle remoto. As duas formas mais comuns de produção de legendas ocultas, ao vivo, são o por estenotipia informatizada e o de reconhecimento de fala. No primeiro método atualmente mais utilizado, os sons são registrados por um estenotipista (através de um estenótipo eletrônico) treinado para digitar em alta velocidade usando um teclado especial que representa letras e grupos de fonemas. O estenotipista registra o que ouve no mesmo momento em que o telespectador assiste ao programa em seu televisor. Este método mais antigo foi criado nos Estados Unidos da América na década de 1970.
O método por reconhecimento de fala já é usado no Brasil com sucesso e grande parte de sua tecnologia foi implementada pelas afiliadas da Rede Globo de Televisão, porém é um recurso muito caro por ser um sistema novo, e nem sempre as emissoras de porte menor não conseguem adquirir e atender estes grupo de telespectadores.

A implantação de um sistema automático para legenda oculta no vídeo consiste em o software de reconhecimento de voz, placa de processamento do texto e encoder para inserção do texto não linhas do vídeo padrão CEA 608 “Linha 21”, norma para televisão analógica e CEA 708 para sistemas digitais. Na figura 3 mostra o fluxo onde o inicia a captura do áudio que é transformada em texto, depois será reprocessada enviada via serial para um equipamento que analisa o padrão do texto e mistura no vídeo, cancelando o original e inserindo o novo produzido.

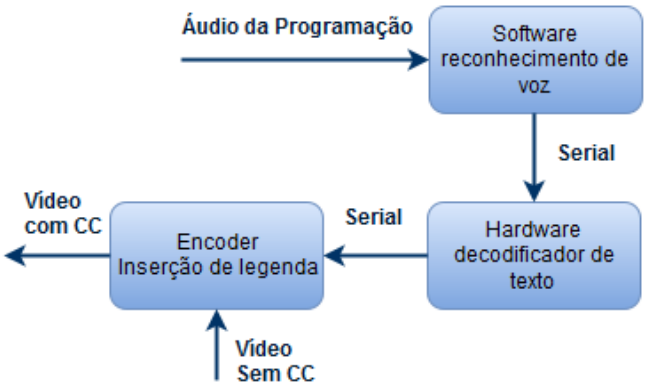

Fig. 3. Sistema para operação automática de closed caption.

Para realizar a inserção da legenda no vídeo é necessário adquirir um equipamento encoder do vídeo, O Link CC Inserter nos padrões CEA-608/708 para SDI (SD e HD) é a solução definitiva para transmissão de CC em redes analógicas e digitais.

Fig. 3. Encoder para inserção de legenda.

V. CONCLUSÕES

Ao capturar o áudio verificado que é muito importante o tratamento do áudio e o armazenamento de banco para treinamento do sistema para uma melhor perfeição, pois cada fala tem uma forma diferente de emissão de som, nesse tipo de sistema é comum o algoritmo reconhecer frases de forma completamente errada, para corrigir tal situação o processo deverá ser submetido a treinamento e referência para um melhor funcionamento.
Apesar do sistema já existir em diversos seguimentos, ainda sim é de extremo interesse que esse projeto siga adiante, pois apenas um áudio como entrada abre diversas possibilidades para criação de aplicações que utiliza uma única placa de controle.
É um grande desafio na confecção do software interpretar corretamente o áudio de entrada, pois não só o reconhecimento do tipo e nível da voz, mas também os ruídos causados pelo som ambiente presente na programação.

REFERÊNCIAS

[1] Ministério Público do Estado do Paraná. Conceitos De Deficiência. Disponível em: < http://www.ppd.mppr.mp.br/modules/conteudo/conteudo.php?conteudo=41> Acesso em: 08 Jun 2016.

[2] CAT, Comitê de Ajudas Técnicas. Comitê de Ajudas Técnicas da área da Pessoa com Deficiência discute reestruturação para ampliar sua atuaçãoDisponível em: < http://www.sdh.gov.br/importacao/2010/10/29-out-2010-comite-de-ajudas-tecnicas-da-area-da-pessoa-com-deficiencia-discute-reestruturacao-para-ampliar-sua-atuacao > Acesso em: 09 Jun 2016.

[3] Um sistema de síntese de voz para a língua Portuguesa Universidade Federal do Rio de Janeiro

[4] Escola de Engenharia Departamento de Eletrônica. pdf

[5] Sistemas de Conversão Texto-Fala: estado da arte, aplicações, arquitectura e desafios Daniela Braga e Miguel Dias. Disponível em: < http://web.letras.up.pt/bhsmaia/EDV/db&md_resumo.htm >.

[6] Como funciona o reconhecimento de voz?. Disponível em: < http://www.tecmundo.com.br/curiosidade/3144-como-funciona-o-reconhecimento-de-voz-.htm> Acesso em: 17 Jun 2016.

[7] Patrick Silva, "Sistemas de Reconhecimento de Voz para o Português Brasileito utilizando os Corpora Spoltech e OGI-22", Trabalho de conclusão de curso, Universidade Federal do Pará, Instituto de Tecnologia, 2008.

[8] Ênio dos Santos Silva, “Desenvolvimento de um reconhecedor automático de voz com suporte a grandes vocabularios para o português brasileiro,” Tech. Rep., 2005.

[9] Rafael Oliveira, Pedro Batista, Nelson Neto, Aldebaro Klautau. "Recursos para Desenvolvimento de Aplicativos com Suporte a Reconhecimento de Voz para Desktop e Sistemas Embarcados". Workshop de Software Livre (WSL), Porto Alegre, 2011. pdf